JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 9, No. 1, June 2026, pp. 177~191
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: https://doi.org/10.36378/jtos.v9i1.5626

Comparison of KNN and Logistic Regression Algorithms in Classifying Food Product Healthiness Based on Nutritional Information

Idris Ajmalul Fikri[1*], Ahmad Homaidi[2], Syarif Aminul Khoiri[3]
[1, 2, 3] Information Technology Study Program, Faculty of Science and Technology, Ibrahimy Sukorejo University, Situbondo, Indonesia

Article Info

Article history:

Received 04 20, 2026
Revised 05 16, 2026
Accepted 06 10, 2026

Keywords:

Classification
K-Nearest Neighbor
Logistic Regression
Nutritional Value
Data Mining

ABSTRACT

Nutritional information on food products can be used to determine the healthiness of a product, but the large number of nutritional attributes often makes it difficult for users to make a quick assessment. This study aims to compare the performance of the K-Nearest Neighbor (KNN) and Logistic Regression algorithms in classifying food product healthiness based on nutritional data. The dataset used comes from Kaggle with a total of 1,204 data points consisting of nine nutritional attributes and one target attribute. The research stages include data preprocessing, normalization using Min-Max Scaling, dividing training and test data, model development, and evaluation using a confusion matrix with accuracy, precision, and recall metrics. The test results show that KNN obtained an accuracy value of 87.67%, a precision of 0.89, and a recall of 0.81. Meanwhile, Logistic Regression obtained an accuracy of 84.93%, a precision of 0.87, and a recall of 0.84. Although Logistic Regression has a slightly higher recall value, KNN shows better overall performance based on accuracy and precision values. Based on the comparative results, KNN was selected as the best model to be implemented in a web-based food product health classification system.

Corresponding Author:

Idris Ajmalul Fikri
Information Technology Study Program
Ibrahimy Sukorejo University
Situbondo, Indonesia
Email: Idrsajmll@gmail.com

1. Introduction

The safety and quality of food products are crucial aspects that play a role in supporting a healthy lifestyle and preventing various diseases related to people's consumption patterns. Choosing the right food not only helps meet daily nutritional needs but also contributes to maintaining health and improving quality of life. Therefore, the ability to identify the healthiness of a food product is becoming increasingly important as part of efforts to raise public awareness of the importance of consuming healthy and nutritious foods [1].

Nutritional information listed on food product labels provides a variety of information about nutritional content, such as energy, fat, saturated fat, carbohydrates, sugars, protein, fiber, and sodium. This information can be used as a basis for assessing the quality of a food product and helping consumers make food choices that meet their health needs. However, the large number of nutritional attributes that must be considered often makes it difficult for consumers to understand and interpret nutritional information

accurately. As a result, the decision-making process regarding food product selection is not always conducted objectively and can potentially result in choices that are less than ideal for nutritional needs [2], [3].

The development of data mining technology enables the processing of large amounts of data to find useful patterns and knowledge [4]. One of the most widely used techniques is classification, which is the process of learning a model based on training data which is then used to predict categories in new data [5]. This approach has proven effective in various domains of food and nutrition analysis. Applying the Random Forest algorithm to classify the composition of processed food menus against toddler nutritional standards yielded 90% accuracy, demonstrating that the data mining approach can produce rapid and objective nutritional compliance assessments [6]. In a similar context, a comparison between Decision Tree and Random Forest algorithms in classifying fast food nutrition showed that Random Forest outperformed Decision Tree with an accuracy of 66.67% versus 55.56%, indicating that ensemble methods are superior in handling the complexity of nutrition data [7]. In addition to label-based classification, machine learning has also been developed for a healthy food recommendation system based on a hybrid approach that combines Content-Based Filtering and K-Means Clustering using Indonesian food nutrition data, with results of precision 0.722, recall 0.740, and NDCG 0.887, demonstrating the great potential of machine learning in supporting healthy food consumption decision making [8]. The results of this study indicate that machine learning methods have great potential in supporting decision-making regarding healthy food consumption.

However, most previous research has focused on using a single algorithm or comparing methods within the same algorithm family. Research specifically comparing distance-based and probability-based algorithms in classifying food product healthiness based on nutritional information is still relatively limited. However, the characteristics of nutritional data, which is dominated by continuous numeric attributes, allow the two approaches to produce different performance in recognizing food product healthiness patterns.

One of the most widely used algorithms in classification is K-Nearest Neighbor (KNN). This algorithm works based on the proximity of data points and determines the class of an object based on the majority of its nearest neighbors [9], Meanwhile, Logistic Regression is a classification algorithm that uses a probability approach to model the relationship between input attributes and target classes [10]. These two algorithms have different characteristics, making them interesting to compare in the problem of food product health classification that uses nutritional attributes as predictor variables.

Based on these conditions, this study aims to compare the performance of the K-Nearest Neighbor and Logistic Regression algorithms in classifying the healthiness of food products based on nutritional information. The dataset used was obtained from the Kaggle platform and contains various nutritional attributes of food products. The performance of both algorithms was evaluated using accuracy, precision, and recall metrics to determine which method provides the best performance. In addition, this study also implemented the classification model into a web-based application so that the research results not only provide academic contributions but also have practical value in supporting the decision-making process related to selecting healthier food products.

2. Research Method

This research uses a data mining approach focused on comparing the performance of classification algorithms. The food product nutritional value dataset was processed using the K-Nearest Neighbor and Logistic Regression algorithms to generate classification models, which were then compared for performance.

To support the implementation and visualization of the results, a web-based system developed using native PHP was used as a tool.

2.1. Literature Review

2.1.1. Nutritional Information

Nutritional Information (NUI) refers to the labeling found on food packaging that provides details regarding nutrient composition and other related information, including serving size, total servings, and the percentage of recommended daily nutritional intake [11]. The information is displayed in numerical form, allowing it to be processed and analyzed computationally.

In this research, nutritional information functions as a dataset composed of numerical attributes that serve as features in the classification stage. The dataset contains several nutritional variables, such as energy, fat, protein, carbohydrates, sugar, and sodium levels. The selected attributes were adjusted to the characteristics of the dataset applied in this study, where each variable possesses different measurement units and value ranges.

Variations in data scale may influence the performance and calculation process of the algorithms. For this reason, a normalization process is applied so that all attributes are transformed into a comparable scale before the classification process is carried out.

2.1.2. Data Mining

Data mining is the process of extracting valuable patterns and information from data sets to support decision-making. One of the main techniques in data mining is classification, a supervised learning method used to build models that map data into predetermined categories. In this study, classification was used to group food products into "Healthy" and "Unhealthy" categories based on numeric nutritional attributes, such as energy, fat, and sugar [12, 13].

2.1.3. Classification

Classification in data mining is a useful process for finding models by analyzing training data sets that provide an overview and differentiate label classes or concepts of data [14], [15]. In the context of this research, classification is used to determine the health category of labeled food products based on nutritional content information such as calories, protein, fat, sugar, and sodium objectively.

2.1.4. K-Nearest Neighbor Algorithm

K-Nearest Neighbor (KNN) is a classification algorithm that determines the category of data based on the number of nearest neighbors in the training data. The classification process is carried out by calculating the distance between the test data and all training data, then the class is determined based on the majority of the data with the closest distance. In this study, the proximity measurement between data uses the Euclidean Distance method because it is suitable for numerical data and can effectively represent similarities between objects [16, 17]. Mathematically, Euclidean Distance is formulated as follows:

$$\mathrm{d}(x, y) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (x_{\mathrm{i}} - y_{\mathrm{i}})^2} \quad (1)$$

Where :
$d(x, y)$: distance between data x and data y
x_i : value of the i-th attribute in the test data
y_i : value of the i-th attribute in the training data
n : number of attributes used

The use of Euclidean Distance in KNN is effective for measuring the similarity of numerical data. To ensure that all attributes contribute equally to the distance calculation process, data normalization is performed before the classification process [18].

2.1.5. Logistic Regression Algorithm

Logistic Regression is a machine learning algorithm used to model the probability of a categorical target variable [19]. Although called a regression method, this algorithm is commonly used for binary classification, such as determining the "Healthy" and "Unhealthy" categories of food products. Unlike K-Nearest Neighbor, which relies on proximity between data points, Logistic Regression uses a probabilistic approach using the sigmoid function to convert the results of a linear combination of attributes into probability values.

The initial stage of the classification process is carried out by calculating the logit (z) value obtained from a linear combination of input attributes and their weights, which is formulated as follows:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (2)$$

Where :
z : Linear combination (logit) results.
β_0 : Constant or bias.
$\beta_1, \beta_2, \ldots, \beta_n$: Coefficients or weights for each nutritional attribute.
$x_1, x_2, \ldots, x_n$: Nutritional content values (such as sugar, fat, sodium).

The z-value is then transformed using the sigmoid function to produce a probability in the range 0 to 1:

$$f(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

Where :
$f(z)$: Probability output (range 0-1).
e : Euler's number or exponential constant (≈ 2.718).

The resulting probabilities are then used to determine the final class based on the threshold value. In binary classification, data with a probability higher than the threshold is categorized into the positive class, while data with a lower probability is categorized into the negative class [20].

2.1.6. Preprocessing Data

This research was designed to compare the K-Nearest Neighbor (KNN) and Logistic Regression algorithms in classifying food product healthiness based on nutritional information. Prior to the classification process, the dataset underwent preprocessing to improve data quality and ensure its suitability for model training and testing.

The preprocessing stage begins with data cleaning, which includes checking for missing values, duplicate data, and data type conformity for each attribute [21]. This process aims to ensure that the data used is free from problems that could affect the classification results.

Next, normalization is performed using the Min-Max Normalization method because each attribute has a different value range. Normalization is necessary to equalize the data scale so that no single attribute dominates the classification process, particularly in the KNN algorithm, which uses distance calculations between data points. The normalization equation used is as follows:

$$x^1 = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

Where :

x^1 = normalized value,
x = original attribute value,
x_{min} = minimum attribute value,
x_{max} = maximum value of the attribute.

Through the normalization process, all attributes have the same scale so that each attribute can provide a more balanced contribution in the process of forming a classification model [22].

After normalization is complete, the dataset is divided into training data and testing data. The training data is used to build KNN and Logistic Regression models, while the testing data is used to evaluate the model's ability to classify previously unprocessed data.

2.1.7. Confusion Matrix

Confusion Matrix is an evaluation technique used to assess the success level of a classification model by comparing the predicted results with actual conditions [23]. This matrix provides a detailed overview of correct and incorrect predictions for each class, making it easier to analyze model performance [22]. This matrix consists of four main components:

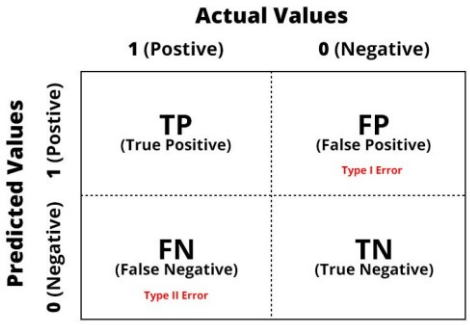

Figure 1. Confusion Matrix

Description:

TP : The amount of data classified as positive that matches the actual condition.
TN : The amount of data classified as negative that corresponds to the actual condition.
FP : The amount of data predicted as positive even though the actual condition is negative.
FN : The amount of data predicted as negative even though the actual condition is positive.

Based on these values, the reliability of the classification system is measured using three main metrics:

Accuracy: Measures the total percentage of correct predictions (both positive and negative) from the entire test data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision: Measures the level of accuracy between the requested data and the predictions provided by the system.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall: Measures the system's success in retrieving information from the positive class in the actual data.

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

In this study, the Confusion Matrix was used as the primary instrument to evaluate and compare the performance of the K-Nearest Neighbor (KNN) and Logistic Regression algorithms. By comparing the accuracy, precision, and recall values of the two algorithms, this study can objectively determine which model is most optimal and consistent in classifying the health status of food products based on nutritional data. [21][10].

2.2. Types of Research

This research is a quantitative study using a data mining approach. The study analyzed numerical data to compare the performance of the K-Nearest Neighbor and Logistic Regression algorithms in classifying food product healthiness based on nutritional value [24].

2.3. Data Sources

The dataset utilized in this research consists of secondary data collected from the Kaggle platform in Comma Separated Values (CSV) format. The data includes nutritional information related to food products, which serves as the main input for the classification process [15, 10].

The data consists of several numeric attributes representing nutritional content, such as energy, protein, fat, carbohydrates, sugar, and sodium. These attributes are used as variables in the data mining process to classify food products' health categories.

2.4. Data Collection Techniques

The data collection technique in this study used the documentation study method [25]. The data used is secondary data obtained from the Kaggle platform through the *Nutrition Dataset for Healthy Food Prediction*. The dataset is available in Comma Separated Values (CSV) format and is used as a data source to compare the K-Nearest Neighbor (KNN) and Logistic Regression algorithms in classifying food product healthiness.

The dataset consists of 1,204 data with 9 feature attributes and 1 target attribute (healthy_label). The attributes used include additives_n, fat_100g, saturated-fat_100g, carbohydrates_100g, sugars_100g, fiber_100g, proteins_100g, sodium_100g, and nutrition-score-uk_100g. Meanwhile, the healthy_label attribute is used as a target variable consisting of two classes, namely healthy (1) and unhealthy (0).

Based on the class distribution, the dataset consists of 686 healthy data sets and 518 unhealthy data sets. This composition indicates a relatively balanced class distribution, minimizing the risk of class imbalance in the classification process. The dataset structure used in this study can be seen in Figure 2.

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10
additives_n	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	sodium_100g	nutrition-score-uk_100g	healthy_label
0.0	50.0	7.81	21.88	3.12	9.4	21.88	0.0	9.0	1
0.0	50.0	6.25	21.88	3.12	6.2	21.88	0.438	12.0	1
0.0	50.0	7.81	21.88	3.12	9.4	21.88	0.0	9.0	1
1.0	0.0	0.0	11.38	0.0	4.1	4.07	0.233	-5.0	1
2.0	2.11	0.53	11.05	8.42	0.0	2.11	0.084	0.0	1
2.0	15.22	6.52	69.57	52.17	0.0	4.35	0.304	24.0	1
1.0	37.5	25.0	50.0	47.5	2.5	5.0	0.1	24.0	0
1.0	36.59	19.51	53.66	51.22	2.4	7.32	0.085	24.0	0
2.0	35.0	20.0	50.0	47.5	7.5	5.0	0.038	21.0	1
1.0	30.0	17.5	55.0	52.5	2.5	5.0	0.15	24.0	1
1.0	17.33	5.33	12.0	0.0	1.3	10.67	2.013	17.0	0
5.0	17.65	8.24	37.65	22.35	1.2	3.53	0.247	17.0	0
2.0	15.38	8.97	74.36	53.85	0.0	2.56	0.256	25.0	0
1.0	78.57	10.71	7.14	0.0	0.0	0.0	0.536	23.0	0
0.0	66.67	6.67	13.33	3.33	6.7	16.67	0.0	9.0	1
1.0	35.71	10.71	0.0	0.0	0.0	57.14	2.857	27.0	0
2.0	20.75	4.25	54.72	30.19	0.9	4.72	0.349	17.0	0
2.0	18.87	3.77	55.66	32.08	0.9	4.72	0.297	17.0	0
3.0	18.87	9.43	27.36	19.81	0.0	4.72	0.132	17.0	0
0.0	1.79	0.45	0.0	0.0	0.0	24.11	0.071	-4.0	1
0.0	30.0	7.5	35.0	15.0	12.5	25.0	0.45	14.0	0
0.0	0.0	0.0	9.17	3.33	2.5	1.67	0.008	-4.0	1
0.0	50.0	12.5	18.75	6.25	6.2	25.0	0.266	15.0	1
0.0	50.0	12.5	18.75	6.25	6.2	25.0	0.266	15.0	1
0.0	0.0	0.0	13.57	9.29	2.9	0.71	0.006999999999999999	-2.0	1
5.0	16.9	7.04	47.89	7.04	1.4	7.04	0.366	15.0	0
0.0	40.0	25.0	37.5	30.0	10.0	7.5	0.012	17.0	1
0.0	40.0	25.0	40.0	30.0	10.0	7.5	0.012	17.0	1
1.0	40.0	25.0	50.0	45.0	5.0	5.0	0.125	22.0	1
6.0	1.25	0.0	0.83	0.42	0.4	0.42	0.075	0.0	1

Figure 2. Nutrition Dataset

2.5. Research Stages

Figure 3. Research Stages

This research was systematically structured to compare the K-Nearest Neighbor and Logistic Regression algorithms in classifying food product healthiness based on nutritional data. The research framework is shown in Figure 3, which illustrates the overall research stages, from data collection to model implementation into the system.

The research stages are as follows:

1. Data Collection

   At this stage, the collection of food product nutrition datasets obtained from the Kaggle platform was carried out as research objects.

2. Data Preprocessing

   This stage includes data cleaning, handling of missing values, and data normalization so that each attribute has a balanced scale.

3. Division of Training Data and Test Data

   The processed dataset is then divided into training data and testing data for the purposes of building and testing the model.

4. Algorithm Implementation

   At this stage, the K-Nearest Neighbor and Logistic Regression algorithms are applied to build a classification model using training data.

5. Model Testing

   The model that has been built is tested using test data to determine the model's ability to classify new data.

6. Model Evaluation

   The test results were then evaluated using a confusion matrix to obtain the accuracy, precision, and recall values of each algorithm.

7. Algorithm Comparison

   This stage is carried out by comparing the performance of the two algorithms based on the evaluation results to determine the method with the best performance.

8. Implementation in the System

   The best algorithm is then implemented into a web-based system using the PHP programming language as a product classification medium.

2.6. System Design

This system design integrates two main functions: as a comparative study tool and as a practical classification instrument. The goal of this design is to explore knowledge gained from the data mining process and then transform it into a practical and useful tool for users in assessing the healthiness of food products.

The system is implemented web-based using the PHP programming language. Data processing is performed through the dataset upload feature and user input of nutritional data. PHP is used to process the data, run the classification algorithm, and display the analysis results directly. With this approach, the system is able to provide information quickly, accurately, and easily understood by users [26].

2.6.1 System Workflow

The designed classification website can be accessed by users by entering the required data or indicators, then the system will process and display the results according to the available functions.

The system workflow designed in this study consists of two main components:

1. Algorithm Comparison

   This section is used to perform the analysis process by comparing two algorithms: K-Nearest Neighbor (KNN) and Logistic Regression. This process includes model training, testing, and performance evaluation using a confusion matrix to determine the best model.

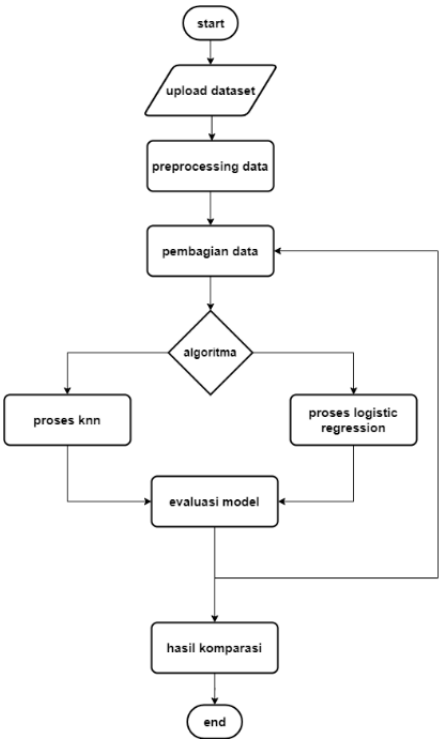

Figure 4. Comparison System Flow

2. Product Classification

The simulation system implements the best model obtained from the comparison process. At this stage, users can test new data to obtain classification results directly.

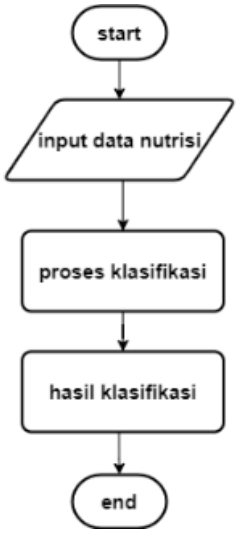

Figure 5. Product Classification System Flow

3. Result and Discussion

A food product health classification system has been successfully implemented as a web-based application using the PHP programming language. This system is designed to compare the K-Nearest Neighbor (KNN) and Logistic Regression algorithms in classifying food product health based on nutritional information. In addition to the comparison, the system also provides a live product classification simulation feature using nutritional data input.

3.1. Test Results

3.1.1. Home Page

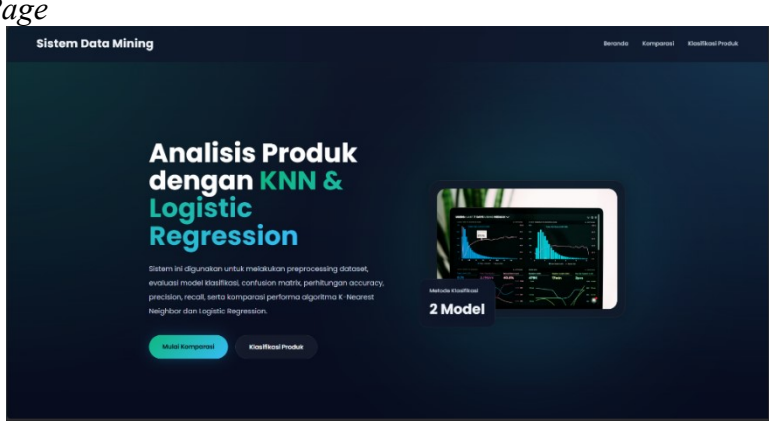

Figure 6. Home Page

Figure 6 shows the main page of the web-based food product classification system. This page features a navigation menu consisting of Home, Comparison, and Product Classification.

The main section of the page displays information about the system's purpose, which is to analyze products using the K-Nearest Neighbor (KNN) and Logistic Regression algorithms. The system also provides two main buttons: the "Start Comparison" button to test the algorithm's performance and the "Product Classification" button to simulate the classification of new nutritional data.

3.1.2. Dataset Upload Page

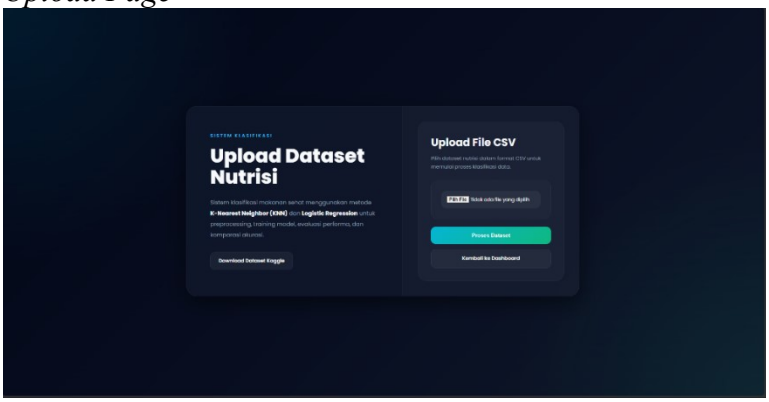

Figure 7. Upload Dataset Page

Figure 7 shows the dataset upload page used to input food product nutritional data in CSV format. This page also provides a Download Kaggle Dataset button that directs users to the research dataset source, allowing for easy and transparent data retrieval. Once the dataset is uploaded, the system will read the data structure and process it in the preprocessing stage before using it in training and testing the K-Nearest Neighbor (KNN) and Logistic Regression algorithms. This page serves as the starting point for the analysis process, as all classification stages depend on the data entered into the system.

3.1.3. Preprocessing Page

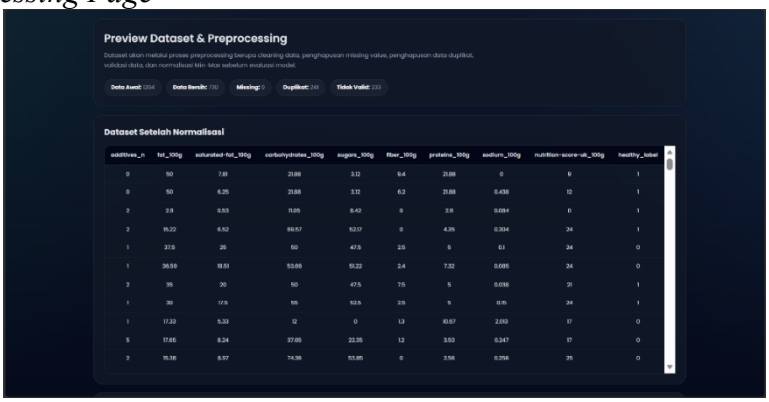

Figure 8. Preprocessing Page

Figure 8 shows the results of the preprocessing stage performed before the classification process. At this stage, the system performs data quality checks and normalization using the Min-Max Scaling method.

The results indicate that the dataset consists of 1,204 data points, with no missing values or invalid data found, and 321 duplicate data points that were handled during the data cleaning process.

After the data cleaning process is complete, all numeric attributes are normalized to the 0–1 range using the Min-Max Scaling method. The normalization results are displayed in tabular form so users can see the changes in the values of each attribute before the data is used in the training and testing of the classification model.

3.1.4. Data Splitting and Model Selection Page

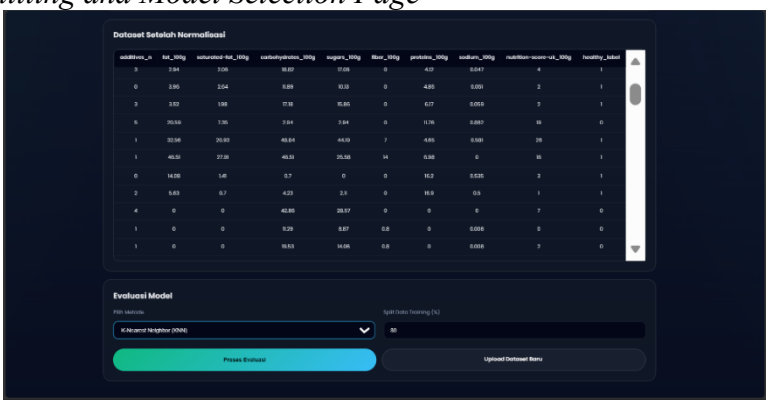

Figure 9. Data Splitting and Model Selection Page

Figure 9 shows the data sharing and classification model selection page. This page displays the normalized dataset, ready for use in model training and testing. Users can select the algorithm to use, whether K-Nearest Neighbor (KNN) or Logistic Regression, and choose the training and testing data sharing ratio.

The system automatically divides the data using a random split method, so that training and test data are randomly selected from the entire dataset. This approach is used to reduce potential bias and provide a more objective model evaluation. After the data split process is complete, the system trains and tests the model using the selected method to generate classification performance evaluation scores.

3.1.5. Model Evaluation Page

The model evaluation page displays the results of testing classification algorithms based on the confusion matrix. At this stage, the system calculates the accuracy, precision, and recall values for each algorithm used.

The evaluation results are displayed in a table and model performance information, allowing users to determine the system's classification success rate.

1. Evaluation Results of the K-Nearest Neighbor Model

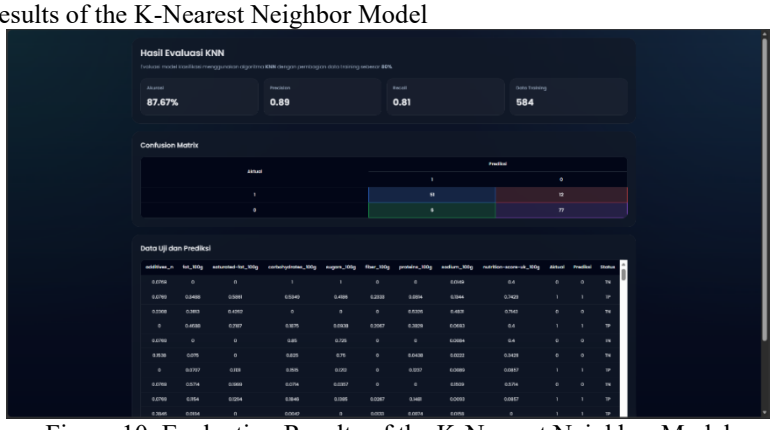

Figure 10. Evaluation Results of the K-Nearest Neighbor Model

Figure 10 shows the evaluation results of the K-Nearest Neighbor (KNN) algorithm. Based on the test results, the KNN algorithm achieved an accuracy of 87.67%, a precision of 0.89, and a recall of 0.81.

The system also displays a confusion matrix, which is used to determine the number of predictions that match and disagree with the actual data. Furthermore, the system displays a table of test data and prediction results, allowing users to directly view the classification results performed by the KNN algorithm on each nutrient data item.

The evaluation results show that the KNN algorithm is capable of classifying the healthiness of food products with a fairly good level of accuracy based on the proximity of nutritional data characteristics.

2. Results of Logistic Regression Model Evaluation

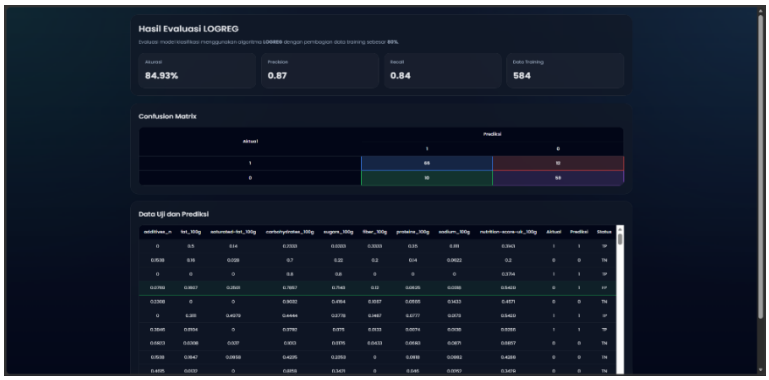

Figure 11. Logistic Regression Model Evaluation Page

Figure 11 shows the evaluation results of the Logistic Regression algorithm. Based on the test results, the Logistic Regression algorithm achieved an accuracy of 84.93%, a precision of 0.87, and a recall of 0.84.

On this page, the system displays a confusion matrix, a table of test data, and classification prediction results. This data is used to assess the model's ability to predict the healthiness of food products based on the nutritional attributes entered.

Based on the evaluation results, the Logistic Regression algorithm has a higher recall than KNN, making it better at recognizing positive data. However, its accuracy is still lower than that of the KNN algorithm.

3. Conditions Before Two Models Run

Figure 12. Before the Two Models Run

Figure 12 shows the system conditions before the evaluation process for both algorithms was run. At this stage, the comparison page cannot display the comparison results because the KNN and Logistic Regression models have not yet been processed by the system.

the system displays a message indicating that the comparison can only be performed after both algorithms have been executed. This ensures that the performance comparison process is based on complete and valid evaluation results.

4. Condition After Two Models Run

Figure 13. After Two Model Runs

Figure 13 shows the system's condition after the KNN and Logistic Regression algorithms were successfully run. After both models were evaluated, the system displayed a comparative performance comparison of the two algorithms based on accuracy, precision, and recall.

3.1.6. Comparison Page

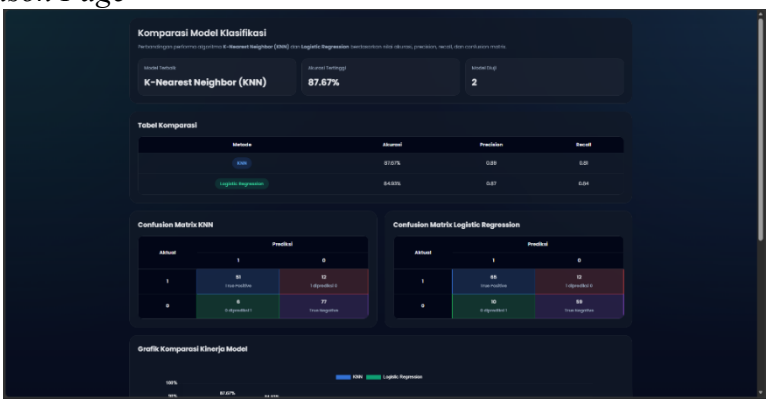

Figure 14. Model Comparison Page

Based on the comparative results, the K-Nearest Neighbor (KNN) algorithm was determined to be the best model in this study. This selection was based on its accuracy value of 87.67% and precision of 0.89, which is higher than Logistic Regression. Although Logistic Regression has a slightly higher recall value, KNN demonstrated better overall performance, thus being selected as the algorithm used in the food product classification feature in the system.

1. Comparison Results

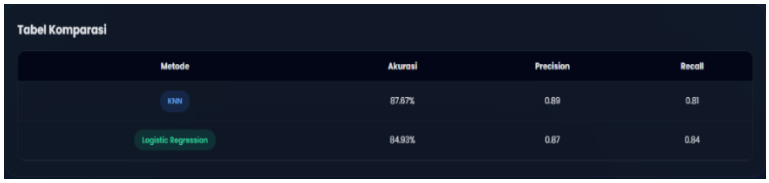

Figure 15. Comparison Result

Based on the Comparison Table, both algorithms performed well in classifying food product healthiness. KNN excelled in accuracy and precision metrics, with scores of 87.67% and 0.89, respectively, while Logistic Regression achieved the highest recall score of 0.84. This difference indicates that KNN is more effective in producing accurate predictions overall, while Logistic Regression has a better ability to identify positive data.

2. Confusion Matrix

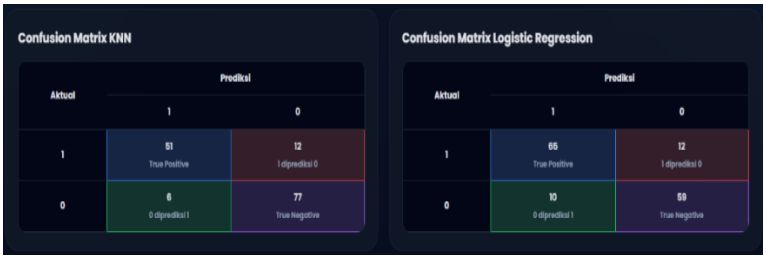

Figure 16. Confusion Matrix

In the KNN confusion matrix, the model successfully classified 51 positive data (true positive) and 77 negative data (true negative) correctly, and only produced 12 misclassifications in the positive class (false negative) and no false positives. These results indicate that KNN can recognize data patterns well and produce relatively accurate predictions. Meanwhile, Logistic Regression also showed good performance with 65 true positives and 59 true negatives, but still produced 10 false positives and 12 false negatives. Although Logistic Regression was able to recognize more positive data, the higher number of false positives caused its accuracy value to be below KNN. Overall, the comparison results show that the KNN method provides better performance because it has a higher level of accuracy and produces fewer misclassifications than Logistic Regression.

3. Comparison Chart

Figure 17. Comparison Chart

Based on the visualization in Figure 17, the performance of the two algorithms is relatively close in all evaluation metrics. However, KNN shows higher values in accuracy and precision, as indicated by the more dominant bar graph compared to Logistic Regression. Meanwhile, Logistic Regression only excels in the recall metric by a small margin. This visualization result confirms the previous evaluation results that KNN has better performance in classifying the healthiness of food products based on nutritional information.

3.1.7. Product Classification Page

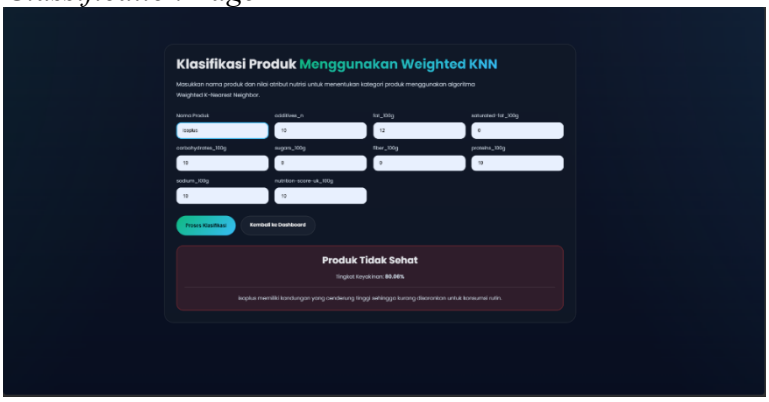

Figure 18. Product Classification Page

The product classification page is used to simulate classification of new nutritional data entered by the user. This page provides an input form containing nutritional attributes such as fat, sugar, protein, sodium, and other attributes based on the research dataset.

The user-inputted data is then processed using the best classification model derived from the algorithm comparison. The system then displays the classification results as "Healthy" or "Unhealthy" based on previously learned patterns from the nutrition dataset.

This page is the primary implementation of the research findings because it allows users to classify food products directly and automatically. Thus, the system can help users understand the healthiness of food products based on the available nutritional information.

3.2. Discussion

Based on the test results, the web-based food product health classification system successfully performed all data processing stages, from preprocessing and model development to evaluation using a confusion matrix, and automated product classification simulation. The system was also able to compare the K-Nearest Neighbor (KNN) and Logistic Regression algorithms using food product nutritional data.

The evaluation results show that the K-Nearest Neighbor (KNN) algorithm achieved an accuracy of 87.67%, a precision of 0.89, and a recall of 0.81. Meanwhile, Logistic Regression achieved an accuracy of 84.93%, a precision of 0.87, and a recall of 0.84. Based on these results, KNN demonstrated better overall performance because it produced higher accuracy and precision values than Logistic Regression.

The superiority of KNN is thought to be influenced by the characteristics of the dataset, which is dominated by continuous numeric attributes, such as fat, sugar, protein, fiber, and sodium content. After normalization using Min-Max Scaling, the Euclidean distance calculation in KNN is able to represent the level of similarity between products more effectively, allowing for better recognition of local patterns in the nutritional data. Furthermore, the relatively balanced class distribution contributes to the stability of KNN's performance in the classification process.

On the other hand, Logistic Regression achieved a slightly higher recall value than KNN. This result indicates that Logistic Regression has a better ability to recognize positive data because it works based on estimated probabilities for each class. However, this increase in recall is accompanied by a decrease in accuracy and precision, indicating that the model produces more classification errors than KNN.

Confusion matrix analysis showed that KNN produced fewer misclassification errors than Logistic Regression. KNN was able to correctly classify most of the healthy and unhealthy data and produced fewer false positives. This indicates that the proximity-based approach is more effective in distinguishing food product characteristics based on their nutritional content patterns than the linear approach used by Logistic Regression.

The results of this study are in line with the research of Junaidi and Meiyanti [9] who stated that KNN is able to produce good classification performance on numerical data because it utilizes the closeness of characteristics between data. This finding also supports the research of Setiawan and Triayudi [26] who showed that KNN has good capabilities in identifying data categories in a web-based classification system. In contrast, although Logistic Regression shows competitive performance, the data proximity-based approach in KNN proved to be more effective for the dataset used in this study.

Based on observations of the test data and predictions of the two algorithms, the attributes nutrition-score-uk_100g, fat_100g, saturated-fat_100g, and sugars_100g showed the most consistent pattern in distinguishing healthy and unhealthy product categories. Products categorized as unhealthy tend to have

higher nutrition-score-uk_100g values, sugar content, total fat, and saturated fat. Conversely, healthy products generally have higher protein and fiber content and lower nutrition-score-uk_100g values.

These findings indicate that attributes related to nutritional quality play a significant role in the classification process. Among all the attributes used, nutrition-score-uk_100g appears to be the most representative indicator because it is a composite score that combines several important nutritional components, such as sugar, sodium, saturated fat, protein, and fiber.

Overall, the research results indicate that a data mining approach can be used as a tool to more objectively evaluate the healthiness of food products. The developed system is capable of simultaneously processing various nutritional attributes and generating automated classification decisions, potentially assisting the public in choosing healthier food products.

Although both algorithms demonstrated good performance, this study still has several limitations. KNN requires distance calculations across the entire training data set, so prediction time can increase with increasing data volume. Logistic Regression, on the other hand, has limitations in handling nonlinear relationships between attributes. Therefore, future research can use larger datasets, implement validation techniques such as k-fold cross-validation, and incorporate other algorithms such as Support Vector Machine (SVM), Random Forest, or XGBoost to obtain more comprehensive comparison results.

Table 1. Changes resulting from research

Condition Before	Method	Condition After
People still have difficulty understanding nutritional information on food products because the data is presented in the form of quite complex numbers.	Development of a web-based classification system using the K-Nearest Neighbor (KNN) and Logistic Regression algorithms.	Users can understand the health categories of food products more easily through the automatic classification results in the system.
There is no media available that can test and compare classification algorithms on food product nutritional data.	Performing data preprocessing, model testing, confusion matrix evaluation, and classification algorithm comparison.	The system is able to display evaluation results in the form of accuracy, precision, and recall so that the algorithm performance analysis process becomes more structured.
The process of determining the health of food products is still done manually and is less objective.	Implementation of data mining methods for health classification of food products based on nutritional value information.	The system is able to automatically determine the classification of food products using the nutritional data provided by the user.

4. Conclusion

Based on the research conducted, a web-based food product health classification system was successfully developed using the K-Nearest Neighbor (KNN) algorithm and Logistic Regression. The system is capable of data preprocessing, model development, evaluation using a confusion matrix, and automatic product classification simulation based on nutritional information.

The comparison results show that the K-Nearest Neighbor algorithm has better performance than Logistic Regression with an accuracy value of 87.67%, a precision of 0.89, and a recall of 0.81. Meanwhile, Logistic Regression obtained an accuracy of 84.93%, a precision of 0.87, and a recall of 0.84. These results indicate that the data proximity-based approach in KNN is more effective in classifying the healthiness of food products than the probabilistic approach used by Logistic Regression on the dataset used in this study.

This research demonstrates that nutritional information can be used as a basis for classifying the healthiness of food products using a data mining approach. The developed system has the potential to help the public understand the healthiness of food products more easily and support more objective decision-making when selecting products based on nutritional information.

Acknowledgement

The author expresses sincere gratitude to Ibrahimy Sukorejo University, particularly the Information Technology Study Program, for the academic support, facilities, and guidance provided throughout the research process and the preparation of this journal. Appreciation is also extended to the supervisor and all individuals who contributed suggestions, motivation, and assistance, enabling this research on the comparison of the K-Nearest Neighbor and Logistic Regression algorithms to be completed successfully.

References

[1] U. K. Yusuf and U. Kalsum, "Hubungan Pengetahuan dengan Tingkat Kepatuhan Membaca Label Pangan Makanan Kemasan pada Mahasiswa Prodi Gizi Universitas Sulawesi Barat," vol. 2023, no. 2, pp. 23–31, 2023.

[2] P. K. Malang, "Analisis label, informasi nilai gizi, kandungan gizi, dan klaim gizi pada produk mp-asi komersial," vol. 14, no. 2, pp. 178–203, 2025.

[3] A. Q. Larasati, Y. I. Putri, T. W. Astuti, R. Putri, E. Mabel, and F. Rohman, "Peningkatan Literasi Gizi melalui Edukasi Label Informasi Nilai Gizi di Posyandu Limau Manis Selatan Kota Padang Strengthening Nutrition Literacy Through Education on Nutrition Labels at Posyandu Limau Manis Selatan , Padang City," vol. 9, no. 1, pp. 86–95, 2026.

[4] S. Pujiono, R. Astuti, and F. M. Basysyar, "Implementasi Data Mining Untuk Menentukan Pola Penjualan Produk Menggunakan Algoritma K-Means Clustering," vol. 8, no. 1, 2024.

[5] R. Obesitas and B. Pola, "Optimasi Metode Random Forest Untuk Klasifikasi Risiko Obesitas Berdasarkan Pola Makan," pp. 56–62, 2025.

[6] D. N. Herisnan and S. Daulay, "Classification of Processed Food Menu Compositions Against Toddler Nutrition Standards Using Random Forest Klasifikasi Komposisi Menu Makanan Olahan Terhadap Standar Gizi Balita Menggunakan Random Forest," vol. 5, no. October, pp. 1498–1507, 2025.

[7] N. I. Yaman, A. R. Juwita, S. Arum, P. Lestari, and S. Faisal, "Perbandingan Kinerja Algoritma Decision Tree dan Random Forest untuk Klasifikasi Nutrisi pada Makanan Cepat Saji," pp. 184–195, 2024, doi: 10.33364/algoritma/v.21-2.1649.

[8] E. Najwa *et al.*, "Penerapan Machine Learning Untuk Rekomendasi Konsumsi Makanan Sehat Berdasarkan Data Gizi Pangan Indonesia," vol. 10, no. 1, pp. 341–348, 2026.

[9] A. Junaidi and R. Meiyanti, "Klasifikasi Status Anak Stunting Menggunakan Metode K- Nearest Neighbor," vol. 10, no. 2, pp. 1435–1445, 2025.

[10] I. R. N. D. Azzahra, Ambarwati, A. Desiani, S. I. Maiyanti, "Jurnal Energy Perbandingan Algoritma K-Nearest Neighbor Dan Logistic," vol. 14, no. 1, pp. 1–8, 2024, doi: 10.51747/energy.v14i1.1843.

[11] N. T. Dewi, L. Yunita, N. Made, W. Sukanty, and F. Ariani, "Edukasi Label Informasi Nilai Gizi sebagai Upaya Peningkatan Pengetahuan dan Kemampuan Membaca Label Gizi Siswa di SMAN 5 Mataram Education on Nutritional Value Information Labels as an Effort to Increase Knowledge and Ability to Read Nutrition Labels for Students at SMAN 5 Mataram," 2023.

[12] A. Hidayat, B. Priyatna, and F. Nurapriani, "Klasterisasi Karakteristik Game Steam Menggunakan Metode K-Means (Studi Kasus : Rilisan Game Tahun 2024)," vol. 9, no. 5, pp. 8890–8894, 2025.

[13] N. A. A. S. Sinaga, B. Saputri, "Pemodelan Classification and Regression Tree (CART) Pada Klasifikasi Gaya Hidup Sehat Menggunakan Pendekatan User-Based Classification," vol. 4, pp. 1028–1036, 2025.

[14] N. B. Putri and A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing Comparative Analysis Of Data Mining Classification Algorithm In Phishing Website Classification," vol. 11, no. 28, pp. 59–66, 2022, doi: 10.34010/komputika.v11i1.4350.

[15] A. V. Agustin and A. Voutama, "Implementasi Data Mining Klasifikasi Penyakit Diabetes Pada Perempuan Menggunakan Naïve Bayes," vol. 7, no. 2, pp. 1002–1007, 2023.

[16] M. I. Mubarok, U. S. Karawang, T. Timur, U. Esai, and N. Indonesia, "Penerapan algoritma k-nearest neighbor (knn) dalam klasifikasi penilaian jawaban ujisan esai," vol. 7, no. 5, pp. 3446–3452, 2023.

[17] S. N. Bakri *et al.*, "Analisis klasifikasi Algoritma K-Nearest Neighboar (K-NN) pada struktur Daerah di Kota Medan," pp. 182–193, 2025.

[18] U. Nijunnihayah and S. S. Hilabi, "Implementation of the K-Nearest Neighbor Algorithm to Predict Sales of Medical Devices in Medical Devices Implementasi Algoritma K-Nearest Neighbor untuk Prediksi Penjualan Alat Kesehatan pada Media Alkes," vol. 4, no. April, pp. 695–701, 2024.

[19] D. A. N. R. Logistik, "Analisis faktor yang mempengaruhi pemilihan gubernur daerah khusus jakarta menggunakan algoritma naive bayes dan regresi logistik 1)," vol. 9, no. 2, pp. 211–224, 2024.

[20] R. Siringoringo, D. Arisandi, E. Kurniawan, and E. B. Nababan, "Model Klasifikasi Dengan Logistic Regression Dan Recursive Classification Model Using Logistic Regression And Recursive," vol. 11, no. 4, 2024, doi: 10.25126/jtiik.1148198.

[21] S. Helmiyah, R. Pramestiawan, and R. Lampung, "Analisis Komparatif Algoritma Machine Learning dengan Metrik Akurasi , Presisi , Recall , dan F1-Score pada Dataset Kacang Kering," vol. 6, no. 3, pp. 152–159, 2025.

[22] F. R. Valerian *et al.*, "Klasifikasi tingkat obesitas menggunakan metode gbm dan confusion matrix," vol. 9, no. 2, pp. 2242–2249, 2025.

[23] A. P. Argadianata *et al.*, "Klasifikasi kualitas buah apel menggunakan metode random forest," vol. 9, no. 2, pp. 2016–2022, 2025.
[24] A. Munthe and B. Ulya, "Penerapan Data Mining Untuk Prediksi Penjualan Roti Terlaris Menggunakan Metode K-Nearest Neighbor," vol. 6, no. 2, pp. 0–7, 2025, doi: 10.47065/bit.v5i2.1783.
[25] I. G. Finalbert, "Prediksi Tren Penggunaan Mobil Listrik Toyota di Washington Menggunakan Model Prophet Berbasis Python," vol. 14, pp. 3650–3657, 2026.
[26] R. Setiawan and A. Triayudi, "Klasifikasi Status Gizi Balita Menggunakan Naïve Bayes dan K-Nearest Neighbor Berbasis Web," vol. 6, no. 2, pp. 777–785, 2022, doi: 10.30865/mib.v6i2.3566.